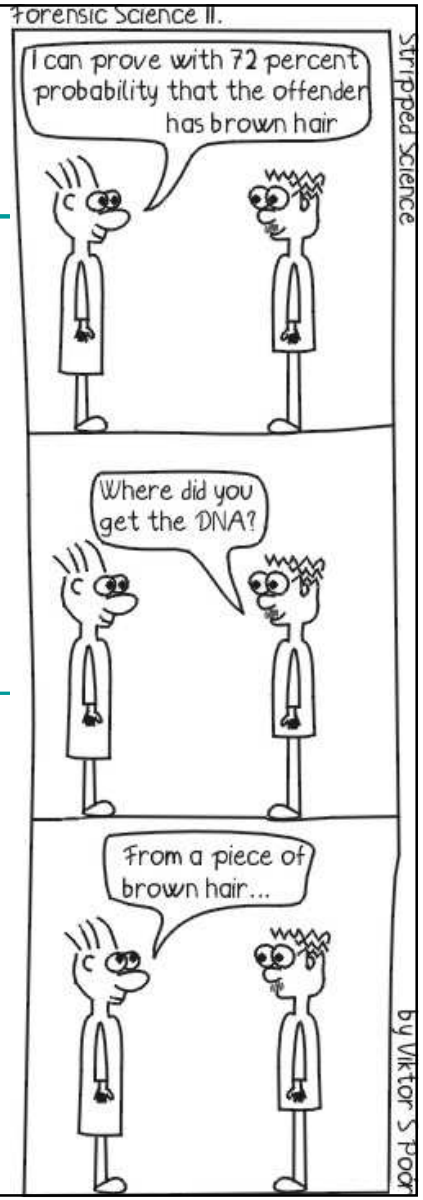
Forensic Automatic Speaker Recognition: Fiction or Science?

Joaquin Gonzalez-Rodriguez
ATVS-Universidad Autonoma de Madrid
http://atvs.ii.uam.es



Motivation

European Network of Forensic Science Institutes

Mission Statement

ENFSI has been established with the purpose of sharing knowledge, exchanging experiences and coming to mutual agreements in the field of forensic science. ENFSI is recognized as an expert group in the field of forensic sciences.

Aim

The aim of ENFSI is to ensure that the quality of development and delivery of forensic science throughout Europe is at the forefront of the world. It will therefore:

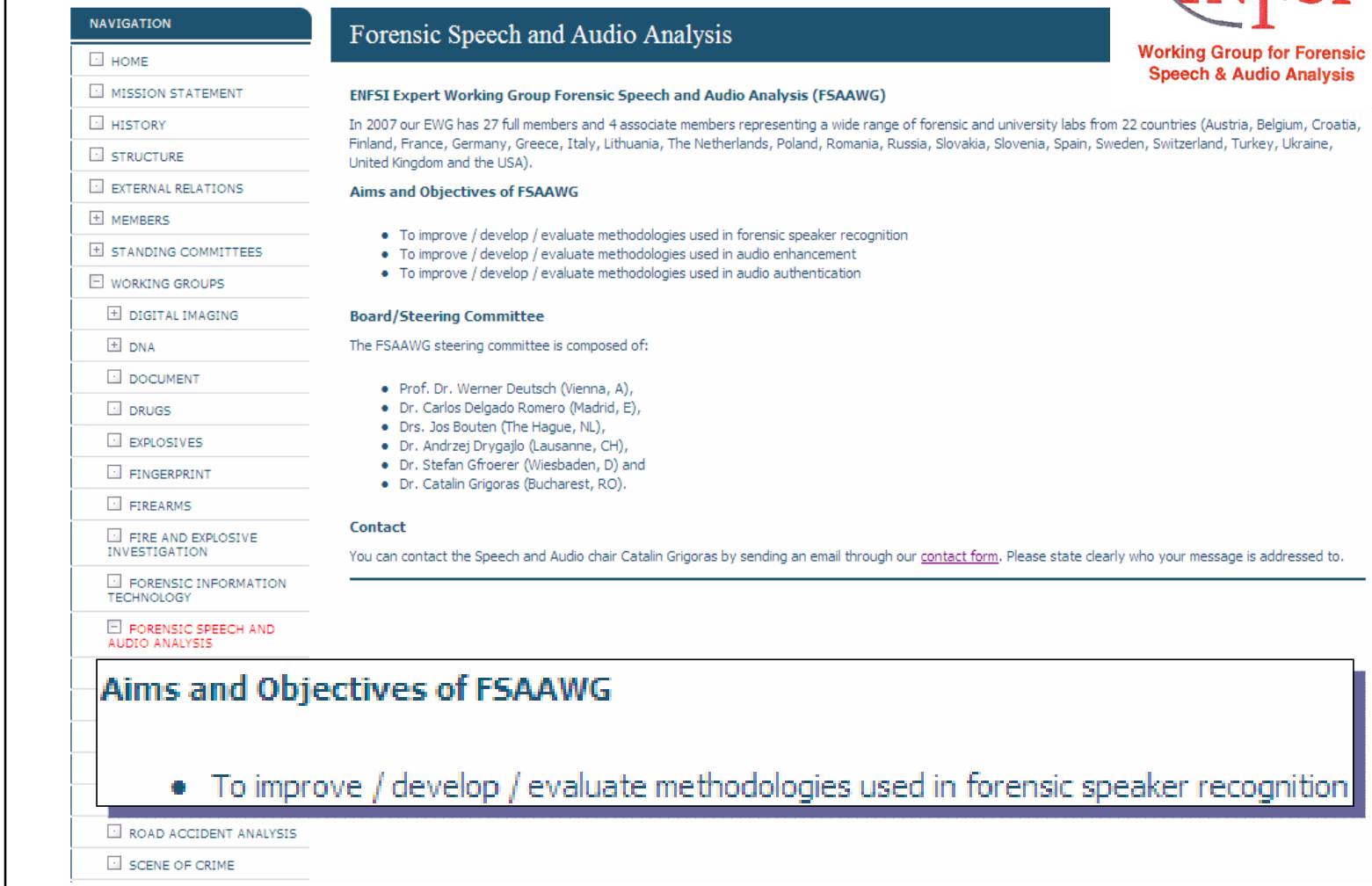
- strengthen and consolidate ENFSI
- expand the membership throughout Europe while maintaining the development and credibility of ENFSI
- establish and maintain a working relationships with other similar organizations
- encourage all ENFSI laboratories to comply with best practice and international standards for quality and competence assurance

Activities

ENFSI activities include:

- organizing meetings and scientific seminars, collaborative studies and proficiency tests
- advising relevant partners on forensic issues
- publishing best practice manuals and glossaries of forensic terms in several languages

1999: FSAA Working Group



FSAAWG Collaborative Exercise

- 2004: NFI (Tina Cambier et al.) prepares a fake case
- Objective: document “methods and reporting strategies”
- Participant labs/experts:
  - 5 auditory-phonetic (all IAFPA members)
  - 5 semi-automatic (variable)
  - 2 fully automatic

- Legend
  - ☑ Correct
  - ? Too short / not enough info / no decision
  - ☒ Wrong

Recording	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Length	120s.	10s.	50s.	50s.	14s.	50s.	5s.	15s.	14s.	18s.
Aud-Phn (3)	☑☑☑	☑☑?	☑☑☑	☑☑☑	☑☑☑	☑☑☑	☑??	☑☑☑	☑☑☑	??☒
SemiAuto (4)	☑☑☑☑	☑???	☑☑☑☑	☑☑☑☑	☑☑??	☑☑☑☑	☑???	☑☑??	☑☑??	☑??☒
FullyAuto (2)	☑?	??	☑☑	☑☑	☑☑	☑☑	??	☑?	☑☑	☑?

Most of the labs report decisions on identification/exclusion or verbal scales of probabilities of identification/exclusion

¿Best practice & international standards?

International Journal of Speech Language and the Law, Vol 14, No 2 (2007)

Current methods in forensic speaker identification: Results of a collaborative exercise

Tina Cambier-Langeveld

Abstract

The field of forensic speaker identification is complicated by the use of a wide array of methods that are employed and it suffers from a lack of evaluation and validation of these different methods. This paper reports on the first internationally oriented collaborative exercise in this field. The main goal of this project was to document the different methods, analyses and reporting strategies that are currently being employed, and to have as many experts as possible participate (representative of the field). Speech material was especially recorded for this purpose and a fake case was constructed. Results from twelve reports are presented anonymously. Furthermore, a detailed account of the most frequently encountered acoustic measurements is given, and a summary of the type and scale of conclusions is provided. These summaries exemplify the huge diversity that seems to govern the field of forensic speaker identification, not only in a global sense but also when reports are viewed in detail.

"… *Reports vary widely on almost every aspect you can think of, and overlap is very limited, also between experts using the same method …*"

Sources of variability in FSR

Photo: http://www.enfsi.eu/page.php?uid=83

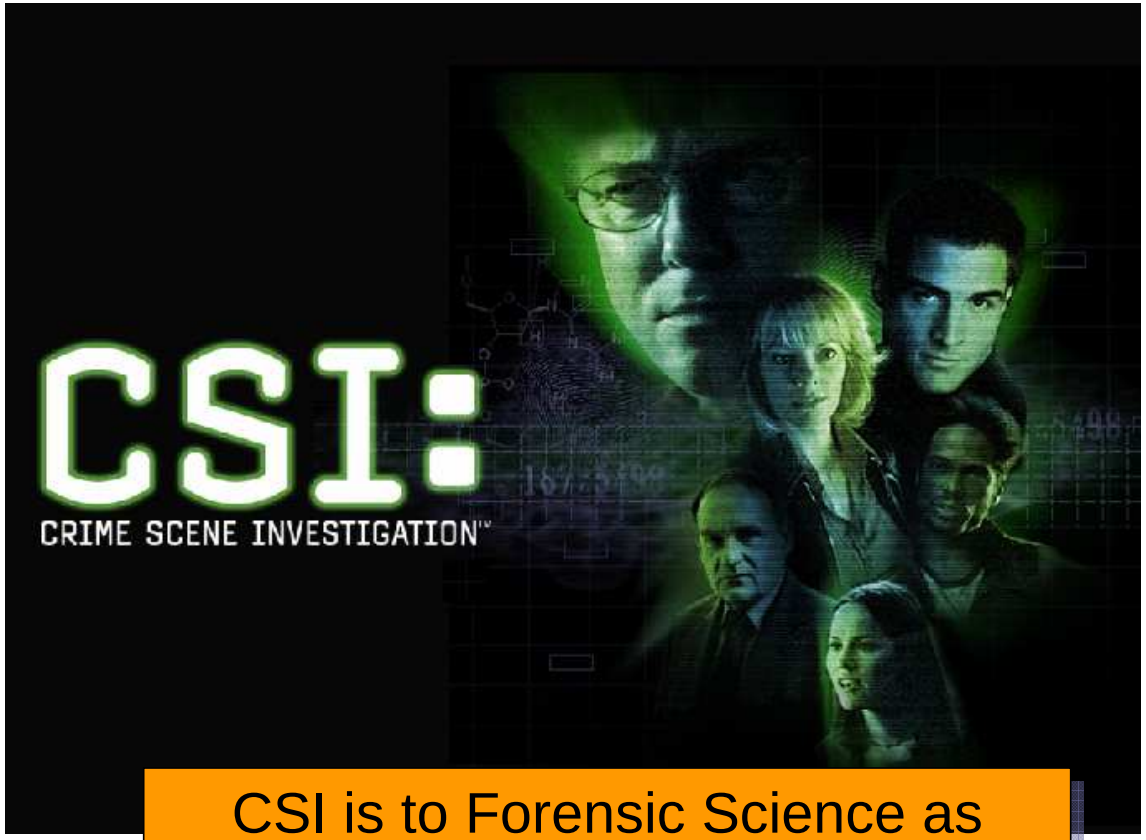
Disparity of:

- Background knowledge: *phoneticians, linguists, engineers, physicists, …*
- Methods: *auditory, acoustic, phonetic, linguistic, semi-automatic, automatic …*
- Tools: *analysis and measurement software, audio equipment, ASR tools, …*
- Reporting: *identification/exclusion, verbal scales of probabilities of identification, …*
- Positions: *crime lab scientists, private practitioners, university staff …*
- Legal systems: *adversarial, inquisitorial*

Quo Vadis, FSR?

- Present FSR shows a combination of two factors:
  - Different methodologies to face the speaker identification problem
  - Influence of “classical” forensic identification
- This talk is:
  - NOT a tutorial on Speaker Recognition
  - NOT a detailed handbook on how to proceed on forensic cases
- We want to learn from the errors and successes of our neighbours:
  - Fingerprint evidence
  - DNA evidence
- **Objective**: to set up a roadmap in order to comply (both trad and auto FSR) with 21st century Forensic Science requirements

What is Forensic Science about?



CSI is to Forensic Science as Science Fiction to Science

Fiction and Science

CSI and its Effects: Media, Juries, and the Burden of Proof

Simon A. Cole
University of California, Irvine - Department of Criminology, Law and Society

The CSI Effect: Popular Fiction About Forensic Science Affects Public Expectations About Real Forensic Science

N.J. Schweitzer
Arizona State University

Michael J. Saks
Arizona State University - College of Law

THE YALE LAW JOURNAL

TOM R. TYLER

Viewing *CSI* and the Threshold of Guilt:
Managing Truth and Justice in Reality and Fiction

CSI Effect

From Wikipedia, the free encyclopedia

It has been suggested that this article or section be merged with *CSI syndrome*. (Discuss)

raising crime victims' and jury members' real-world expectations of forensic science

evidence in court.

Courts and Forensic Science

- *"Judges and lawyers usually react to science with all the enthusiasm of a child about to get a tetanus shot. They know it's painful and believe it's necessary, but haven't the foggiest idea how or why it works."*

  Black et al.: "Science and the Law After *Daubert*" *Texas Law Review* 1994.

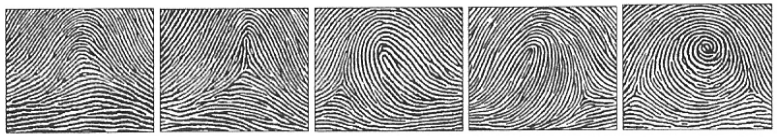
Forensic Identification Sciences

Fingerprinting

From C. Champod et al., Fingerprints and Other Ridge Skin Impressions, CRC Press 2004

Illustrations of the Three Levels of Friction Ridge Skin Features

Level	Illustration
Level 1	Mairs's family tree is illustrated in Figure 2.2; basic general patterns are illustrated below:

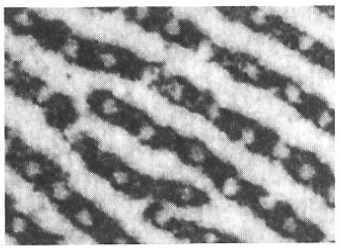


Simple arch Tented arch Right loop Left loop Whorl

Level 2 Ridge endings, bifurcations, and dots are the basic minutiæ (ridges are shown in black); all other types are combinations of these:

Ridge ending Bifurcation Dot

Level 3



Inked impression showing succession of pores and specific shapes of edges (friction ridges are in black ink)

Fingerprint Reporting

- Based in its high discrimination power, three possible states for reporting:
  - *Identification*: detection of more than N minutiae (N~12-16)
  - *Exclusion*: clear differences
  - *Inconclusive*: detection of less than N minutiae
- For decades considered "*the golden standard of forensic identification*"
- Fingerprint experts have long claimed:
  - "*Absolute certainty of identifications and zero error rate*"
  - *"Probable, possible, or likely identification are outside the acceptable limits of the science of friction ridge identification*", (SWG-FAST 2002)

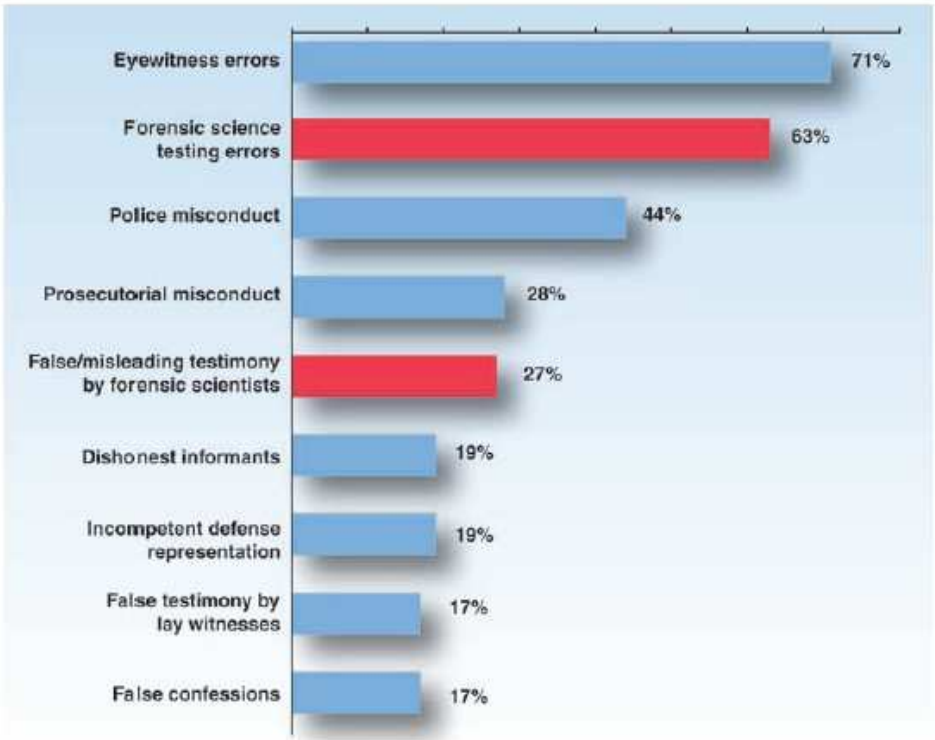
Forensic Identification Reporting

- All identification-of-the-source areas use solid analytical procedures:
  - Chemical analysis
  - Firearms
  - Toolmarks & Shoemarks
  - Fibers
  - Voice (acoustic, phonetic, linguistic, signal processing, pattern recognition)
- Highly influenced by fingerprinting, once a set of observations is obtained:
  - The expert (subjectively)
    - weighs the similarities and dissimilarities
    - set thresholds for comparison

  between questioned and control samples to produce a conclusion
- Conclusions are reported as
  - One of three states: Identification / Exclusion / Inconclusive
  - Verbal scale of probability of identification (M levels)
    - E.g., the suspect is likely/very likely/extremely likely to be the author


INNOCENCE PROJECT
SEARCH
Quick Links
KNOW THE CASES : UNDERSTAND THE CAUSES : FIX THE SYSTEM
ABOUT : DONATE : NEWS & RESOURCES
Rickey Johnson
Served 25 years in Louisiana for a crime he didn't commit.
September 6, 2008 : 220 EXONERATED
Español ►
June 22, 2006
180 EXONERATED
The World
POLITICAL
>200 wrongly convicted in US
x 195 countries = ????
© 2001 Maps.com - http://www.maps.com

Forensic Id Science & Conviction Errors



prints, prints. core iences o sup- ed but taking exper- tist or ne tra- regard d, and

aliza- n: that been foren- e evi- to the (7, 8). tion of his as- fferent Thus, rvably

npe, AZ School 78712,

Fig. 1 Factors associated with wrongful conviction in 86 DNA exoneration cases, based on case analysis data provided by the Innocence Project, Cardozo School of Law (New York, NY), and computed by us. Percentages exceed 100% because more than one factor was found in many cases. Red bars indicate factors related to forensic science.

5 AUGUST 2005 VOL 309 SCIENCE www.sciencemag.org

Major errors

U. S. Department of Justice
Office of the Inspector General

A Review of the FBI's Handling of the Brandon Mayfield Case

UNCLASSIFIED AND REDACTED

Office of the Inspector General
Oversight and Review Division
March 2006

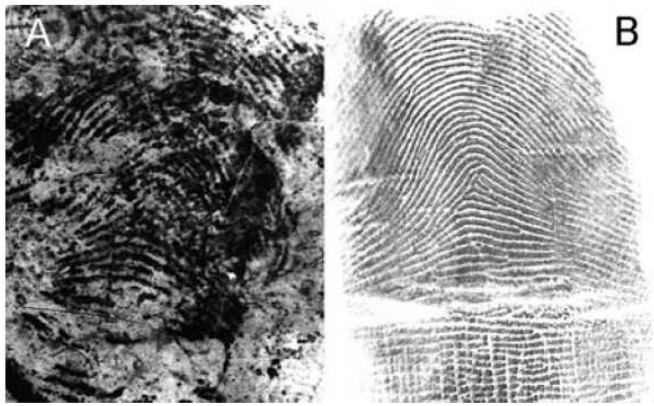


Fig. 4. (**A**) A latent fingerprint believed to belong to a terrorist involved in train bombings in Madrid, Spain, in March 2004. (**B**) A database print belonging to Brandon Mayfield of Portland, Oregon. On the basis of these prints (though not necessarily these very images), FBI fingerprint examiners erroneously identified Mayfield as the bomber (*26*). [Source: Problem Idents, onin.com/fp/problemidents.html#madrid]


Home > Special Reports
Chicago Tribune
chicagotribune.com
FORENSICS U
Guilty, said bite expe
MILWAUKEE — In a cubic
former prosecutor are tryin
forensic discipline. More
Exonerated by DNA,
In the fall of 2002, DNA tes
8-year-old girl in Billings, M
prison. More
Judge doubts lip pri
A Kane County judge has
CNN INTERNATIONAL .com
U.S. Edition
Member Center: Sign In | Register
MAKE CNN.com YOUR HOME PAGE
Every week on CNN International
Watch online now
world's untold stories
SEARCH
THE WEB
CNN.com
SEARCH
CNN.com Home Page -
More CNN.com sections:
ON CNN TV
SCHEDULE
Schedule | Anchors & Reporters | Contact us |
CNN Programs
CNN PRESENTS
CNN Presents airs on Saturdays and Sundays at 8 p.m.
CNN Student News
Educator's Guide for CNN Presents
CONTACT US
Send your questions and comments to CNN Presents.
EVIDENCE
Reasonable Doubt
Can Crime Labs be Trusted?
A joint investigation by CNN and the Center for Investigative Reporting examines the lack of standards, quality controls and training at many of the nation's forensic laboratories and raises serious doubts about some forensic scientists.
GALLERY: Key cases and labs
QUIZ: Test your knowledge
ARCHIVE: More from CNN Presents
Airs: May 6 at 6 a.m. and 3 p.m. and
May 7 at 6 a.m. and 2 p.m. ET.
ATVS

EDITORIAL

Forensic Science: Oxymoron?

In detective novels and television series, criminals often get caught because they leave fingerprints at the scene. Well, art does imitate life; fingerprint analysis is widely used in U.S. courts and those of many other countries. But last year a funny thing happened to fingerprint evidence on the way to a conviction. Applying the standard set for the admissibility of scientific evidence by the U.S. Supreme Court in the 1993 Daubert case, Judge Louis Pollak ruled that an expert could not testify that the prints at a crime scene matched those of a suspect. Shock reverberated through the criminal justice community, until Judge Pollak induced a sigh of relief from district attorneys everywhere by saying that at least in this case, such testimony could be used after all.

The Supreme Court's Daubert standard has generated some ambiguity for the legal community, but the Court did list several criteria for qualifying expert testimony: peer review, error rate, adequate testing, regular standards and techniques, and general acceptance. Judge Pollak's initial finding was that the evidence flunked all but one. Some distinguished legal scholars think that he was right on that call and wrong on the second. The resulting controversy has reignited some old challenges to "forensic science."

It's not that fingerprint analysis is unreliable. The problem, rather, is that its reliability is unverified either by statistical models of fingerprint variation or by consistent data on error rates. Nor does the problem with forensic methods end there. The use of hair samples in identification and the analysis of bullet markings exemplify kinds of "scientific" evidence whose reliability may be exaggerated when presented to a jury. Some criminal defense attorneys have become concerned about the degree to which processing and enhancement of such images could mislead jurors who believe they are seeing undoctored originals. PhotoShop, after all, is everywhere.

ionic interests—security and justice—would be furthered by a more scientific and reliable
r analyzing crimes. The mystery here is why the practitioners don't seem to want it!

Donald Kennedy
Editor-in-Chief

ce has remained firm

www.sciencemag.org SCIENCE VOL 302 5 DECEMBER 2003

UAM UNIVERSIDAD AUTONOMA DE MADRID

REVIEW

The Coming Paradigm Shift in Forensic Identification Science

Michael J. Saks[1] and Jonathan J. Koehler[2]

Converging legal and scientific forces are pushing the traditional forensic identification sciences toward fundamental change. The assumption of discernible uniqueness that resides at the core of these fields is weakened by evidence of errors in proficiency testing and in actual cases. Changes in the law pertaining to the admissibility of expert evidence in court, together with the emergence of DNA typing as a model for a scientifically defensible approach to questions of shared identity, are driving the older forensic sciences toward a new scientific paradigm.

Little more than a decade ago, forensic individualization scientists compared pairs of marks (handwriting, fingerprints, tool marks, hair, tire marks, bite marks, etc.), intuited whether the marks matched, and testified different, criminalists conclude that the marks were made by the same person or object.

Although lacking theoretical or empirical foundations, the assumption of discernible uniqueness offers important practical benefits that different objects share a common set of observable attributes. Without the discernible uniqueness assumption, far more scientific work would be needed, and criminalists would need to offer more tempered opinions in court.

Legal and scientific forces are converging to drive an emerging skepticism about the claims of the traditional forensic individualization sciences. As a result, these sciences are moving toward a new scientific paradigm. [We use the notion of paradigm shift not as a literal application of Thomas Kuhn's concept (*9*), but as a metaphor highlighting the transformation involved in moving from a pre-

New paradigm (1/2): admissibility

- Admission of evidence:
  - Relevance (to the case)
    - Exceptions: non-evidental constraints (time, resources), illegally collected
  - Competence (of the expert) ⇨ difficult for judges
    - How the “expert” obtains his/her conclusions from observations is not questioned !!!
- US Supreme Court (Daubert, 1993): expert testimony must be both:
  - Relevant
  - Reliable: **conclusions** derived from the scientific method
- “General guidelines” can be summarized in:
  - **Testability**: accuracy/reliability, proficiency testing, data supported
  - **Transparency**: clear & detailed reporting, replicability, standards, motivation of each step of the analysis

US Federal Rules of Evidence (before 2000)

ARTICLE VII. OPINIONS AND EXPERT TESTIMONY

Rule 702. Testimony by Experts

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise

US Federal Rules of Evidence (from 2000)

ARTICLE VII. OPINIONS AND EXPERT TESTIMONY

Rule 702. Testimony by Experts

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

- Daubert criteria & FRoE.702
  - Apply to US Federal Courts
  - Sets the highest standard to be fulfilled
    - ⇨ likely to be followed by others (countries?, courts …)
      - *"… specially with voice …",* Hodgson, 2007.

New paradigm (2/2): DNA Profiling

- DNA analysis has become the new *"golden standard"* in Forensic Identification Science:
  - Scientifically based
  - Avoids experience-based opinions
  - Clear and standard procedures
  - Probabilistic, avoiding hard "match" or "non-match" statements
  - *Two-factor approach* to assess the weight of the evidence:
    - *Similarity* factor
    - *Typicality* (or *rarity*) factor

Likelihood Ratio approach as model of clear, standard and probabilistic framework

Bayesian inference of identity: the likelihood ratio approach

Forensic casework

- Two exclusive hypothesis:
  - Prosecution hypothesis, H_p : *the suspect is at the origin of the recovered samples*
  - Defense hypothesis, H_d : *a different person (unknown) is at the origin of the recovered samples*
- Evidence, E :
  - comparisons between recovered and suspect samples
- Information of the case, I :
  - police investigations, witness and victims testimonies, etc.
- A priori probabilities: $P(H_p \mid I)$, $P(H_d \mid I)$
  - Derived from I
  - Unknown to the scientist, independent of E

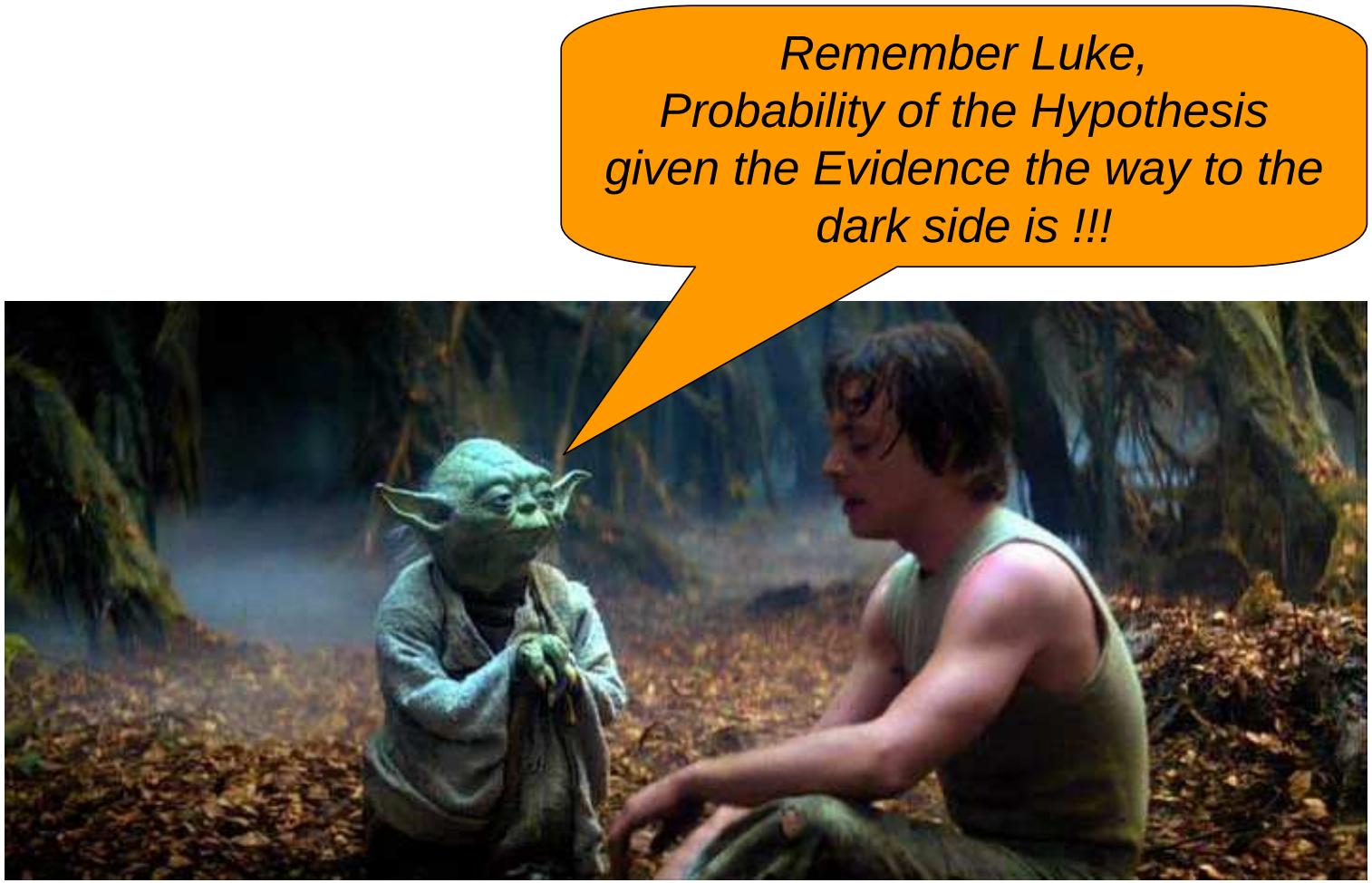
The Court question

- All interested parties (Court, Police …) want to know:
  - *How probable is that the suspect said the incriminating speech, given the evidence adduced in support?*

$$¿P\left(H_p \middle| E, I\right)?$$

- *Example:*
  - *a cow is suspect of having eaten the garden grass. Given a witness observed that the offender has four legs (E), what is the probability of the offender being a cow (H_p)?*
- The scientist CAN NOT quote this probability !!!
- Moreover, we would be ignoring the (unknown) prior probabilities
  - E.g., very high similarity but defendant proves succesfully an alibi

Take home message !!!



The Forensic Scientist Role

- The Forensic Scientist CAN ONLY quote, with the observed evidence E:
  - $P(E \mid H_p, I)$ ⇨ similarity
    - e.g. if a "match", $P(E \mid H_p, I) = 1$
    - Within-source (intra-) variability
  - $P(E \mid H_d, I)$ ⇨ typicality
    - e.g. random match probability
    - Between-source (inter-) variability

Reasoning with probabilities

- Example: a forensic scientist reports $P(E \mid H_d, I)$
  - *"The probability of the observed similarities with the suspect voice, given that the questioned recording comes from an innocent person, is 1 in 100"*.
- Prosecution interpretation:
  - Then, the suspect is GUILTY with probability

    (1 – 1/100) = 0.99 = 99%
- Defense interpretation:
  - As we know the criminal is an adult male from Madrid (~1.000.000), there are 10.000 (1%) possible authors. Then, the suspect is INNOCENT with probability

    (1 - 1/10.000) = 0.9999 = 99.99%

Interpretation fallacies

- Prosecution fallacy:
  - Error in transposing the conditional probability

$$P(H_p \mid E, I) \neq 1 - P(E \mid H_d, I)$$

- Defence fallacy
  - Logically correct
  - Fallacy: not all adults male in Madrid are equally likely than the suspect (I)
  - If the suspect comes from a database search, OK!
- Reporting probabilities is NOT a recommended practice
  - Judges and juries can be easily misled!

The odds form of Bayes theorem

$$P\left(H_p \middle| E, I\right) = \frac{P\left(E \middle| H_p, I\right) P\left(H_p \middle| I\right)}{P\left(E \middle| I\right)}$$

$$P\left(H_d \middle| E, I\right) = \frac{P\left(E \middle| H_d, I\right) P\left(H_d \middle| I\right)}{P\left(E \middle| I\right)}$$

$$\frac{P\left(H_p \middle| E, I\right)}{P\left(H_d \middle| E, I\right)} = \frac{P\left(E \middle| H_p, I\right)}{P\left(E \middle| H_d, I\right)} \times \frac{P\left(H_p \middle| I\right)}{P\left(H_d \middle| I\right)}$$

Posterior odds (O_{post}) = LR (Likelihood Ratio) X Prior odds (O_{prior})

LR : scientist role

O_{prior} : court role

Moreover: ¿$P\left(H_p \middle| E, I\right)? = \frac{LR \cdot O_{prior}}{1 + LR \cdot O_{prior}}$

Role of the forensic scientist

- Estimation of the *likelihood ratio*

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)}$$

similarity

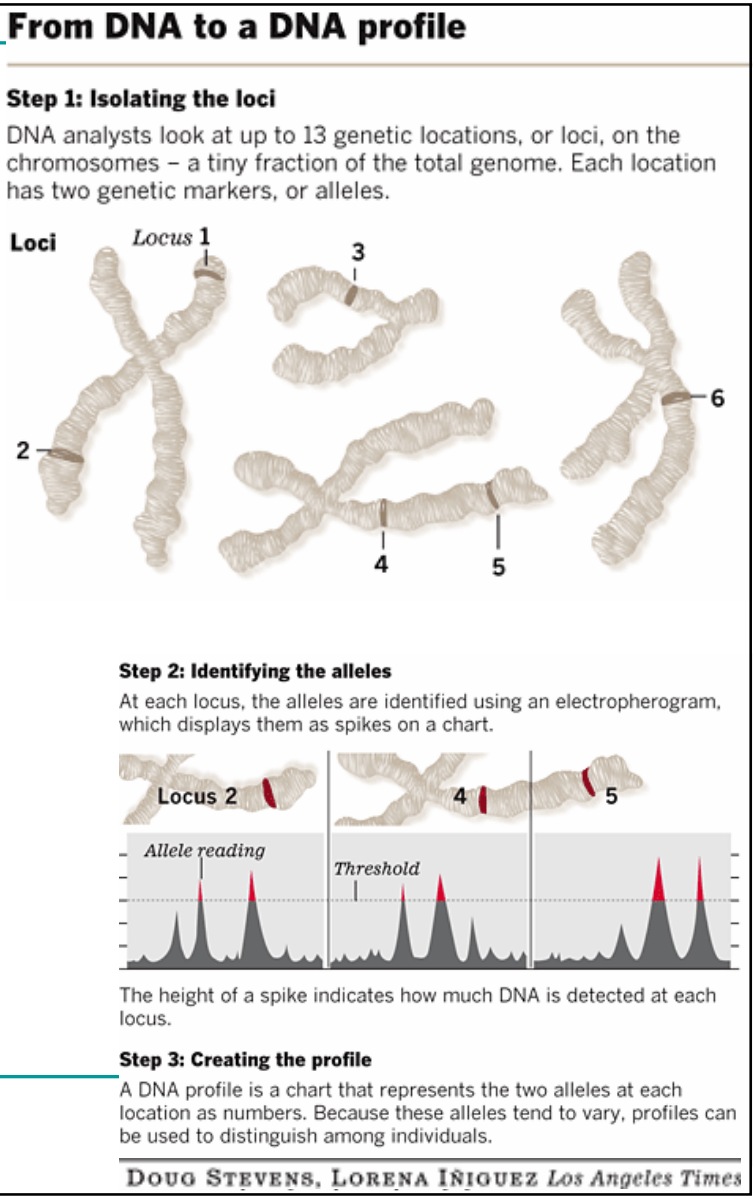
typicality

- The bigger (smaller) than one the LR value, the stronger the support to the prosecution (defense) hypothesis

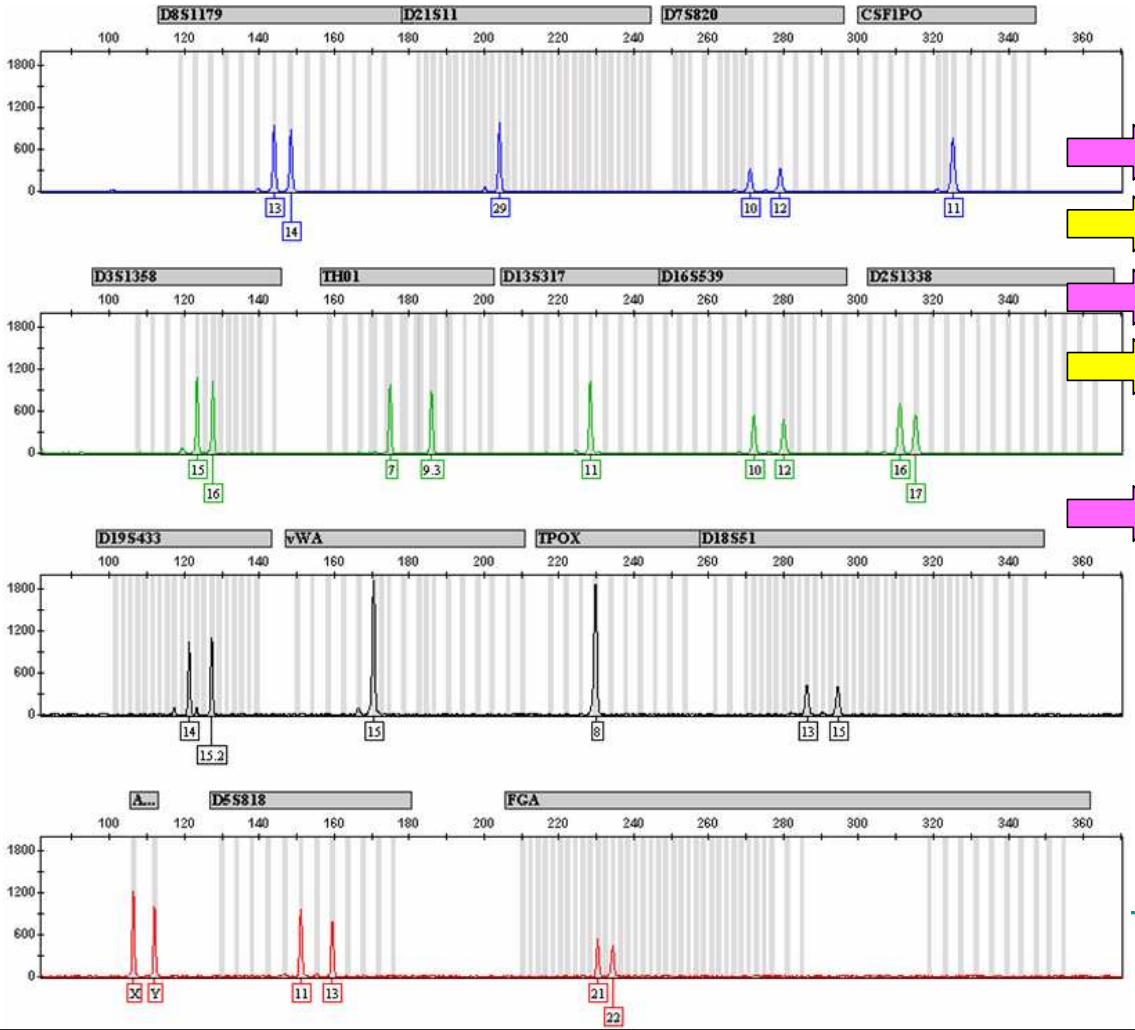
Discrete and Continous Likelihood Ratio estimation

DNA Profiling

- DNA contains genetic instructions to encode the different biological functions
- Non-coding parts (98%) contain at different locations (*loci*) highly variable number of repetitive sequences of nucleotides called *Short Tandem Repeats* (STR)
- At each *locus*: two specific numbers (*alleles*) of repetitions of the given sequence of nucleotides
  - Inherited from father & mother
- STRs are stable within individuals but vary greatly between individuals



Sample 16 loci DNA Profile



Profile (16 loci):

D8S1179 (13,14)

D21S11 (29,29)

D7S820 (10,12)

CSF1PO (11,11)

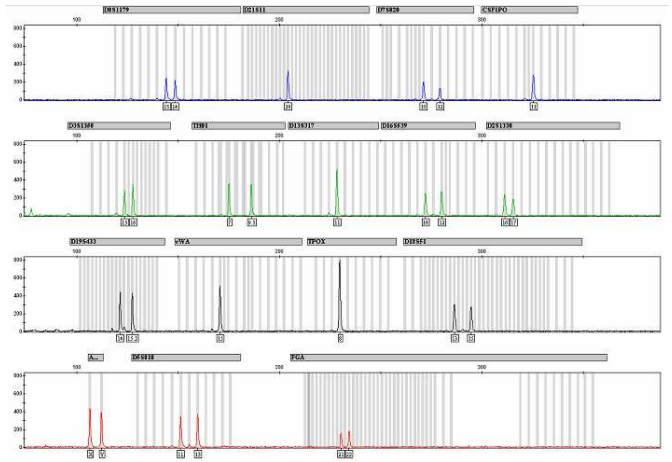
….

FGA (21,22)

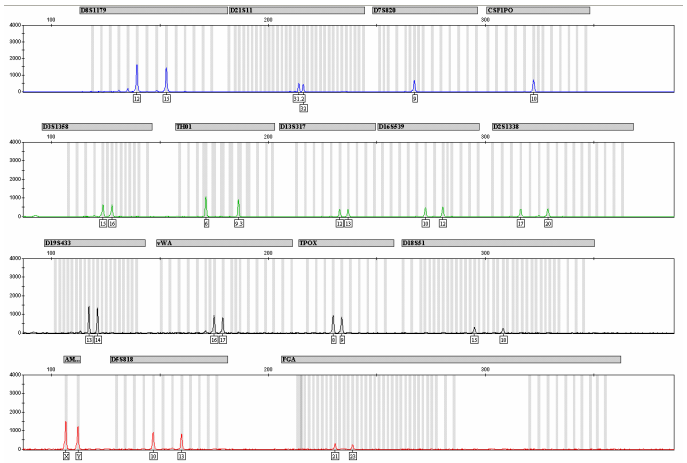
Homozygous (single peak locus)

Heterozygous (two peaks locus)

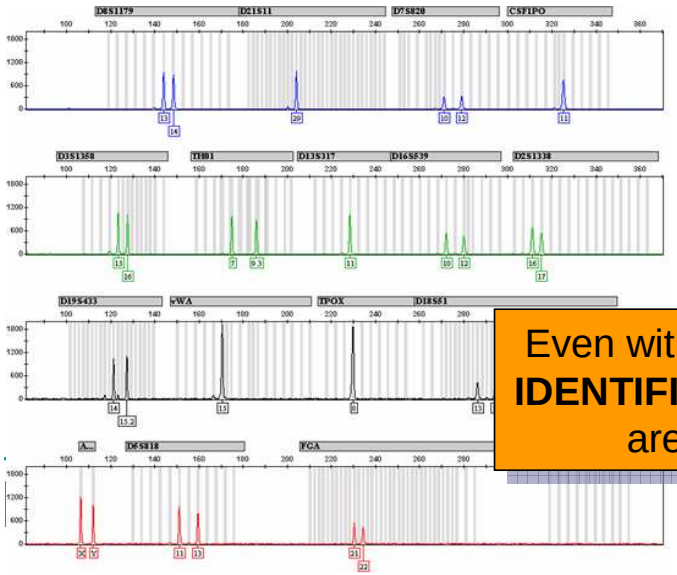
UNIVERSIDAD AUTÓNOMA DE MADRID



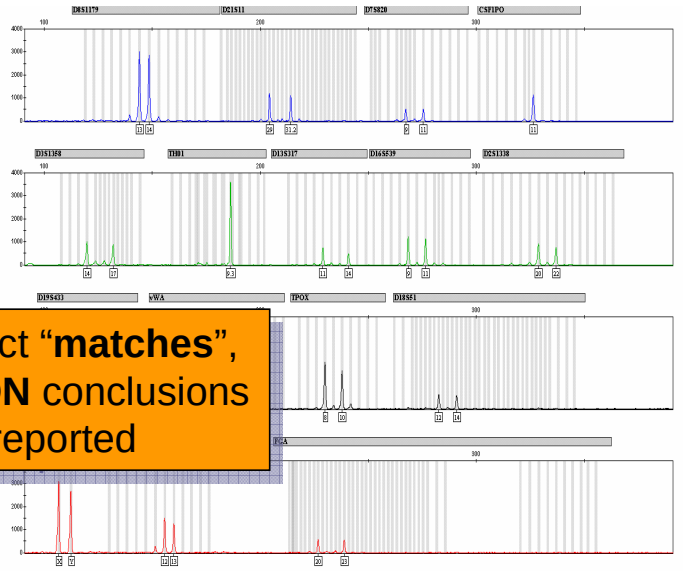
Matching profiles



Non-matching profiles

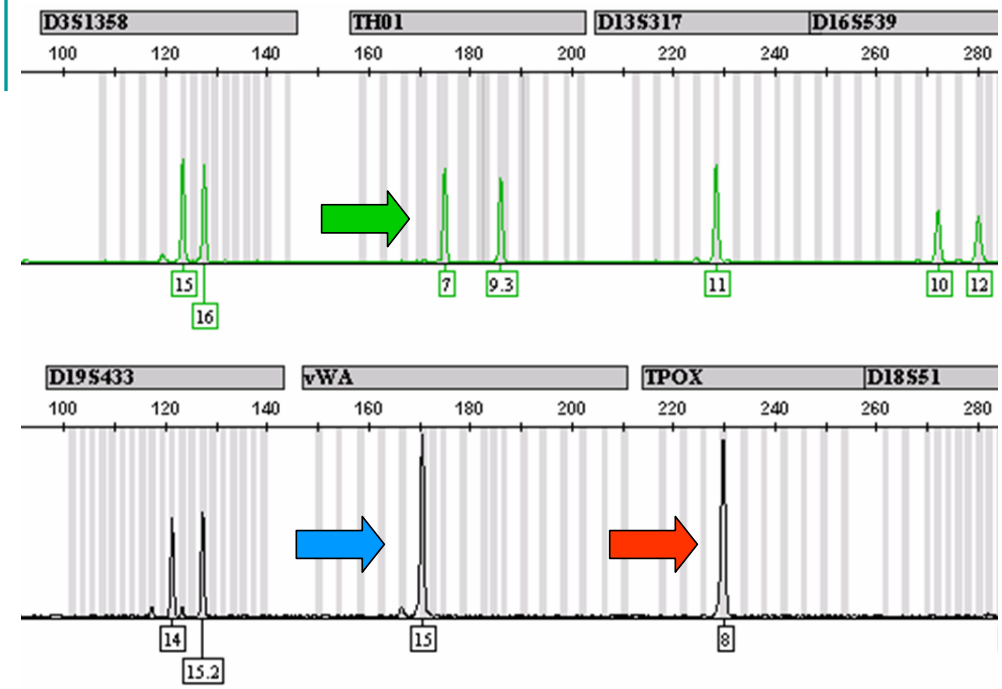


Even with perfect "**matches**", **IDENTIFICATION** conclusions are **NOT** reported



Probability of a DNA profile

- Linkage equilibrium *(between-loci)*:
  - Alleles appearing on one locus are independent of the alleles appearing on any other locus
- Hardy-Weinberg equilibrium *(within-locus)*
  - Each allele on a locus appears independently of each other allele on that locus
- Pr_i ⇨ probability of allele i in a given population
- Probability for a genotype (allele pair):
  - Homozygous: $Pr_{ii} = Pr_i \; x \; Pr_i$
  - Heterozygous: $Pr_{ij} = 2 \; x \; Pr_i \; x \; Pr_j$



Locus	Allele	Relative Frequency
TPOX	6	0.006
	8	0.506
	9	0.094
	10	0.063
	11	0.294
	12	0.038
VWA	14	0.086
	15	0.067
	16	0.276
	17	0.300
	18	0.205
	19	0.048
	20	0.001
	21	0.001
THO1	6	0.295
	7	0.147
	8	0.184
	9	0.232
	9.3	0.026
	10	0.116

Table from D. Lucy, *Introduction to Statistics for Forensic Scientists*, Wiley, 2005

Hardy-Weinberg:

- THO1 (7, 9.3) = 2 x 0.147 x 0.026 = 7.644×10^{-3}
- VWA (15, 15) = 0.067 x 0.067 = 4.489×10^{-3}
- TPOX (8, 8) = 0.506 x 0.506 = 0.256036

Frequency of the 3 locus profile:

8.7856×10^{-6}

(linkage equilibrium):

ATVS

UNIVERSIDAD AUTÓNOMA DE MADRID

Discrete LR estimation: DNA

- Pattern of the suspect matches the one at crime scene
- Assuming uncontaminated samples, no relatives involved, error free operational procedures:
  - Probability of a match given H_p

$$P(E|H_p, I)=1$$

  - How frequent is that pattern in the relevant population:

$$P(E|H_d, I)= 8,7856 \cdot 10^{-6}$$

  - The Likelihood Ratio is (3 loci):

$$LR = 113.822,6$$

Typical LR values (16 loci) ~ billions !!!

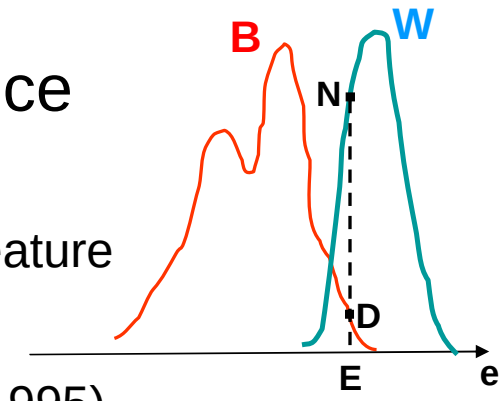
Even then, they do not report “identification” but LR or RMP !!!

Continous LR estimation

$$LR = \left. \frac{f\left(e \middle| H_p, I\right)}{f\left(e \middle| H_d, I\right)} \right|_{e=E}$$

- Numerator: from suspect samples
  - Within-source variability (W)
- Denominator: from relevant reference population
  - Between-source variability (B)

- Types of evidence
  - *e* real valued
    - Score / single feature
  - *e* feature vector
    - MVLR (Aitken, 1995)



LR=N/D

Assessment of Forensic Automatic Speaker Recognition Systems

NIST Speaker Recognition Evaluations

- NIST SRE's have become a *de facto* standard in ASR
- New data is recorded and released through LDC
- Variety of
  - Speaking conditions: conversational & interview (2008)
  - Channel conditions: telephone, mobile, multiple mics
  - Train/test lengths & sessions
- Participants submit both a *score* (real number) and a *decision* (T/F) per speech eval pair
  - e.g., ~ 50.000 trials (~3.600 target and ~47.800 non-target) from ~600 spkrs in main eval condition (1c1c) at SRE06

2006 Evaluation Participation

- 36 submitting sites

Australia	Canada	China (6)
Czech Republic	Denmark	Finland
France (8)	Germany (2)	Israel
Italy	Lebanon	Singapore (2)
South Africa	Spain (2)	Switzerland
United Kingdom	United States (6)	

- 96 systems
- 283 test condition/system combinations

Fig. 2. Participation in the 2006 evaluation.

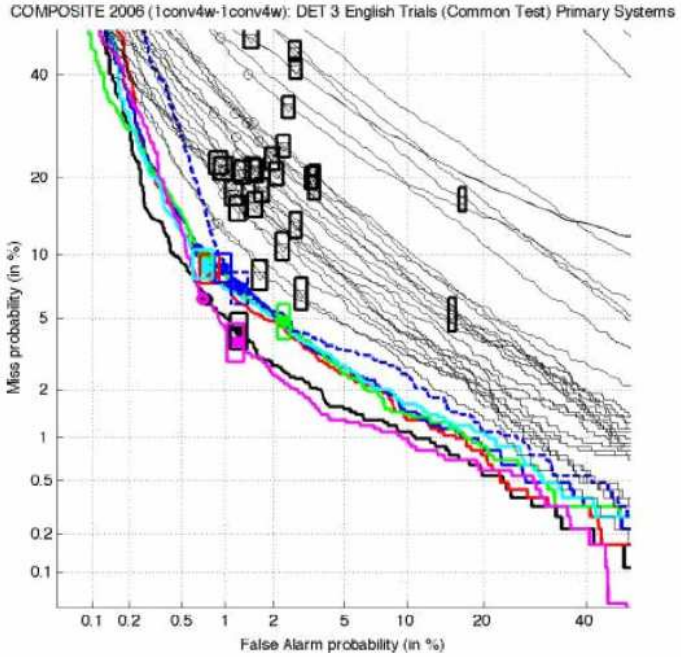


Fig. 3. Common condition results for 36 primary systems in the 2006 evaluation.

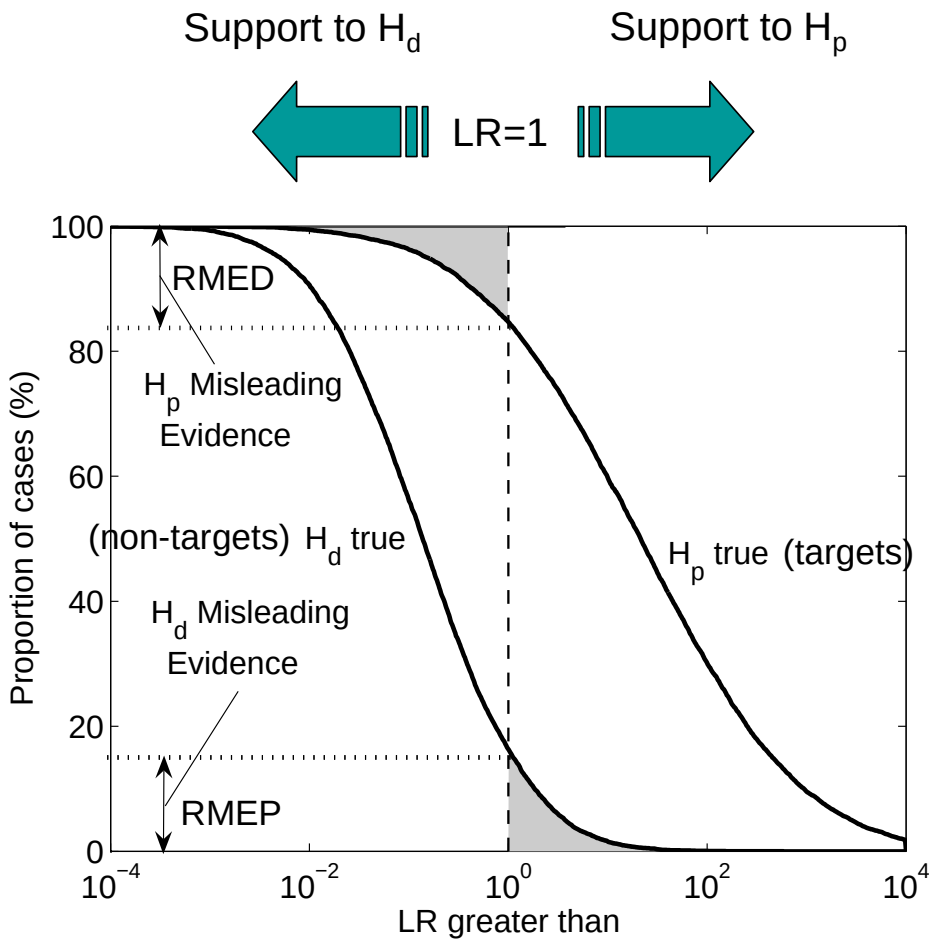
V. Participants

Participation in the NIST SREs has been growing over the years. There were 24 participating sites in both 2004 and 2005, more than in any previous evaluations. Then, in 2006, the number of participants grew by 50%, with 36 different sites (or teams of sites) submitting one or more evaluation systems. The participating sites included research labs from companies, nonprofit org… America, Eur… reflecting wo… core test DET …nts in the 2006 …highlighted …

DET plots are a good measure of discrimination

Without a threshold (court!), scores have NO meaning

Assessment of Forensic LR values: Tippet plots



- Two (1-cpd(LR)) curves when H_p or H_d are true
- Discrimination is shown as separation between curves
- Ideal system:
  - H_p true curve > LR=1
  - H_d true curve < LR=1
- RMEP/RMED
  - Rate of misleading evidence in favour of the prosecution/defense

UAM UNIVERSIDAD AUTONOMA DE MADRID

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computer Speech and Language 20 (2006) 230–275
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Application-independent evaluation of speaker detection

Niko Brümmer [a,b,*], Johan du Preez [b]

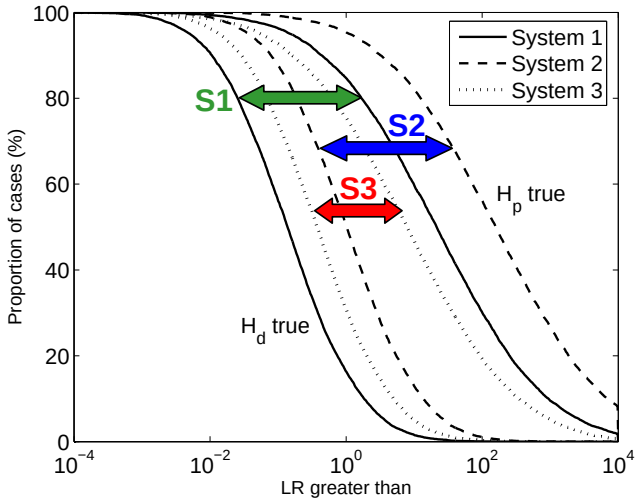
[a] *Spescom DataVoice, Stellenbosch 7600, South Africa*
[b] *University of Stellenbosch DSP Group, South Africa*

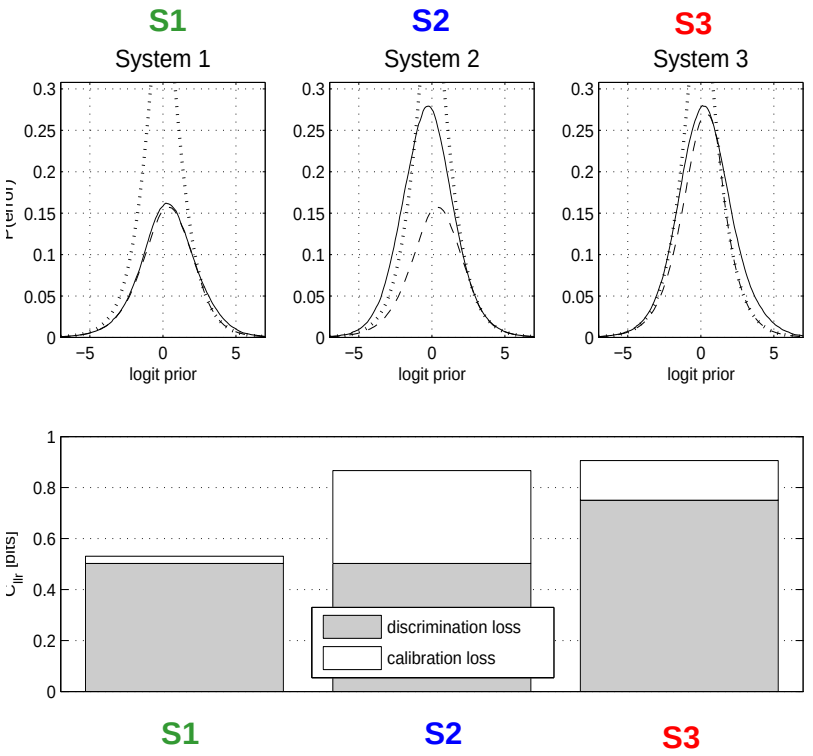
Received 1 November 2004; received in revised form 1 August 2005; accepted 3 August 2005
Available online 6 September 2005

We propose and motivate an alternative to the traditional error-based or cost-based evaluation metrics for the goodness of speaker detection performance. The metric that we propose is an information-theoretic one, which measures the effective amount of information that the speaker detector delivers to the user. We show that this metric is appropriate for the evaluation of what we call *application-independent* detectors, which output soft decisions in the form of log-likelihood-ratios, rather than hard decisions. The proposed metric is constructed via analysis and generalization of cost-based evaluation metrics. This construction forms an interpretation of this metric as an expected cost, or as a total error-rate, over a range of different application-types. We further show how the metric can be decomposed into a *discrimination* and a *calibration* component. We conclude with an experimental demonstration of the proposed technique to evaluate three speaker detection systems submitted to the NIST 2004 Speaker Recognition Evaluation.

application-types. We further show how the metric can be decomposed into a *discrimination* and a *calibration* component. We conclude with an experimental demonstration of the proposed technique to evaluate

Effects of miscalibration: an example





S2 = S1 + offset ⇨ both have exactly the same DET

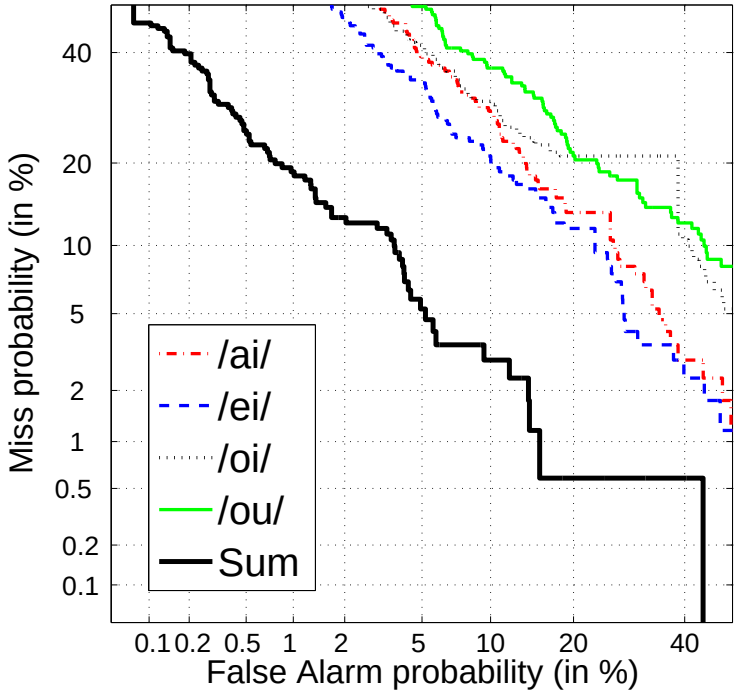
Discrimination is not enough !!!

Low calibration loss is a must !

LR estimation from speech evidence

"Trad" LRs: DET assessment

- LRs derived from formant frequencies in Australian diphtongs



Diphtongs APE plots

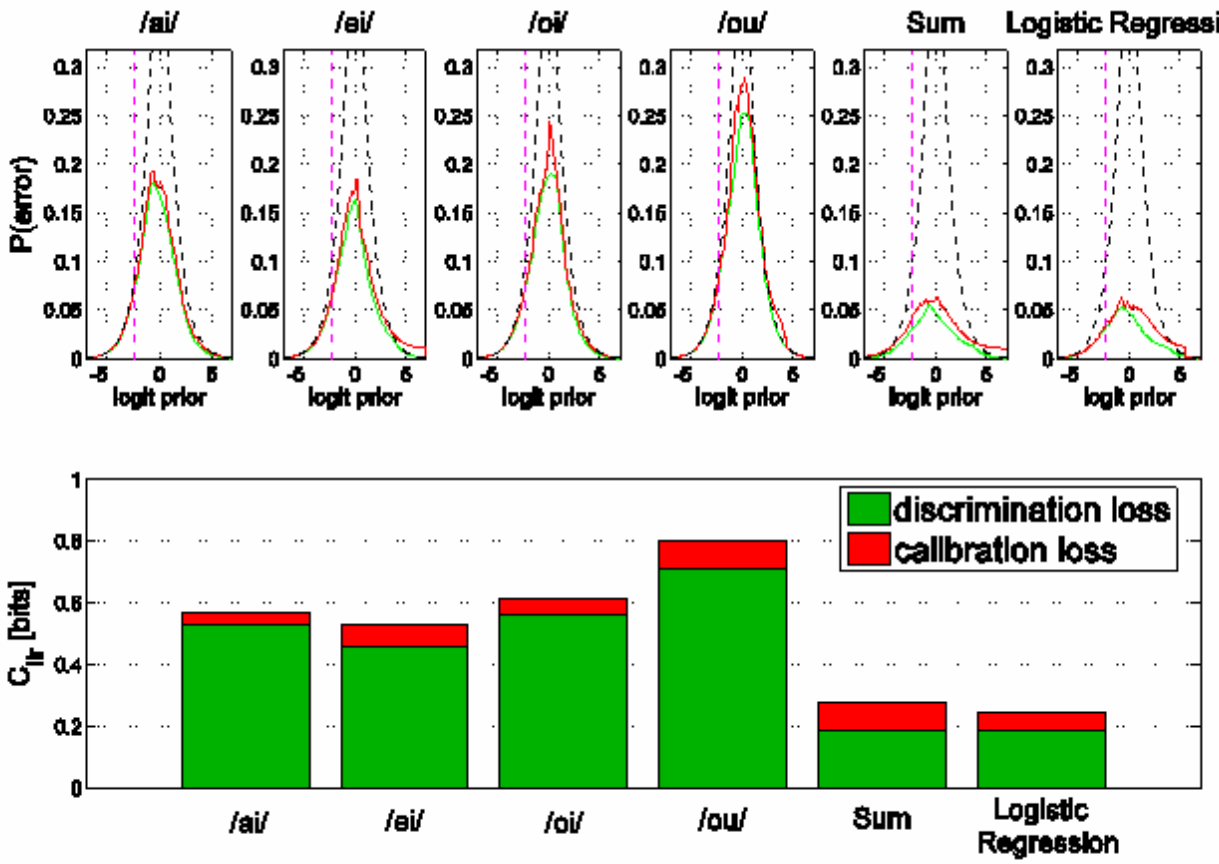


Fig. 3. APE and C_{llr} plots for the four-diphthong discrimination.

"Auto" LRs in NIST SRE'08

- Two types of test speech:
  - Phonecall conversational speech (Mixer 3)
    - Phonecall-phn: telephone recording
    - Phonecall-mic: simultaneous multiple microphone recording
  - Interview speech (Mixer 5)
    - Interview-mic: multiple simultaneous microphone recording

Training channel	Test channel	# trials
Telephone	Telephone	24128
	Microphone	8746
Microphone	Telephone	6693
	Microphone	19776
	Total male	**59343**

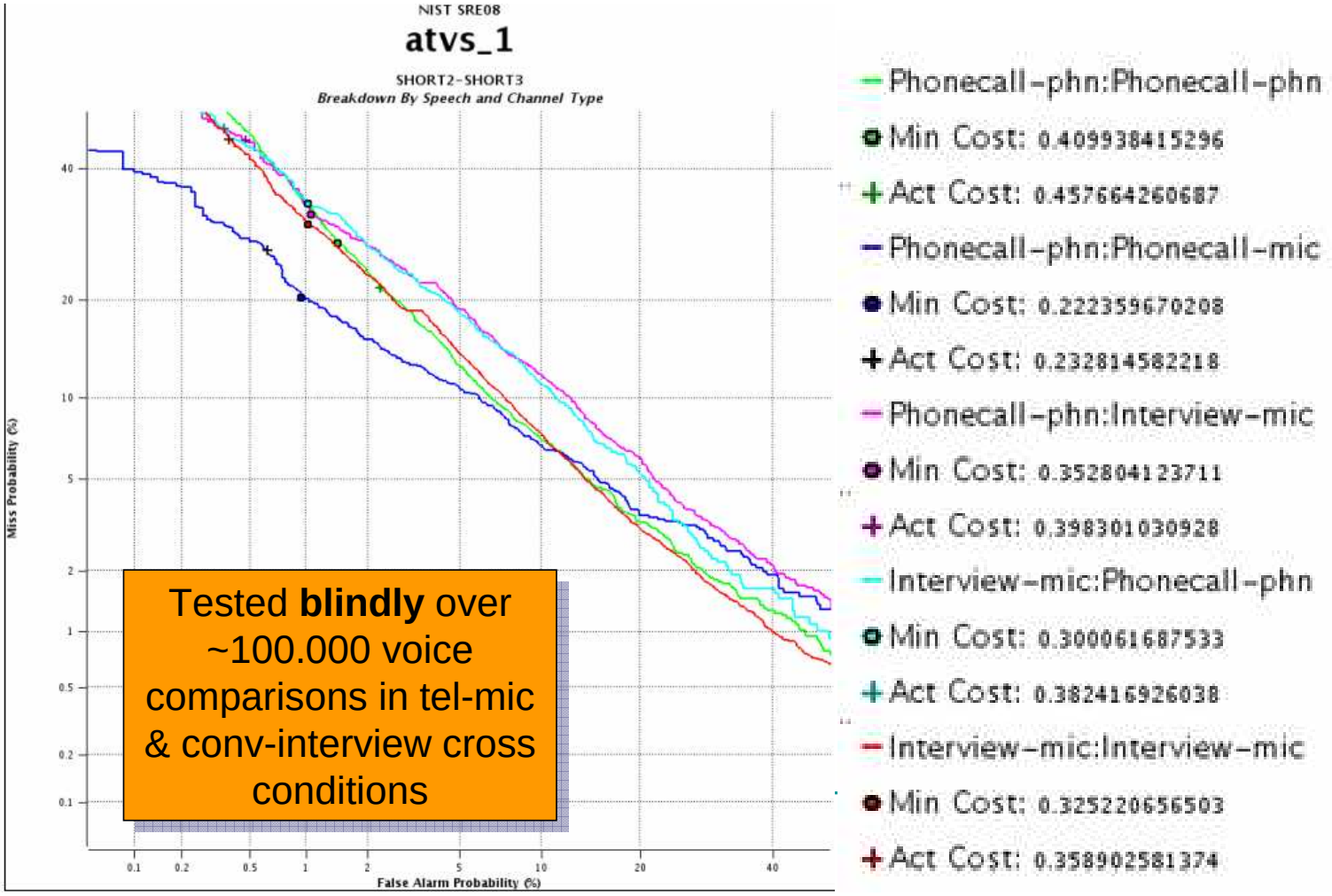
Training channel	Test channel	# trials
Telephone	Telephone	12922
	Microphone	7025
Microphone	Telephone	5048
	Microphone	14270
	Total female	**39265**

1788 Mixer 3 (conversational) spk models

1475 Mixer 5 (interview) spk models

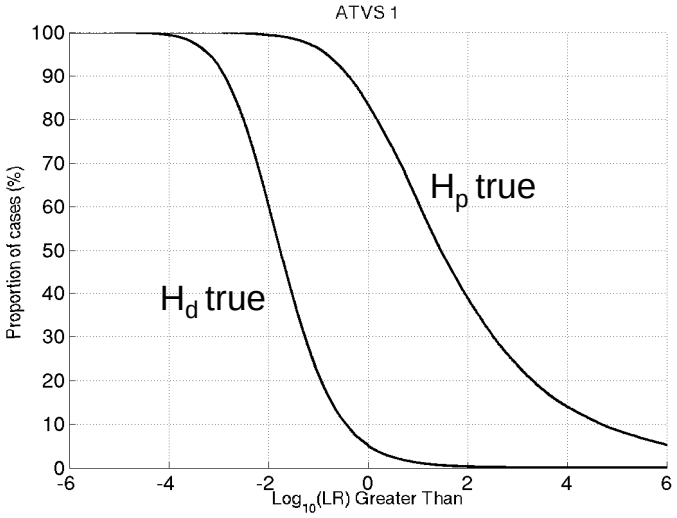
UAM UNIVERSIDAD AUTONOMA DE MADRID

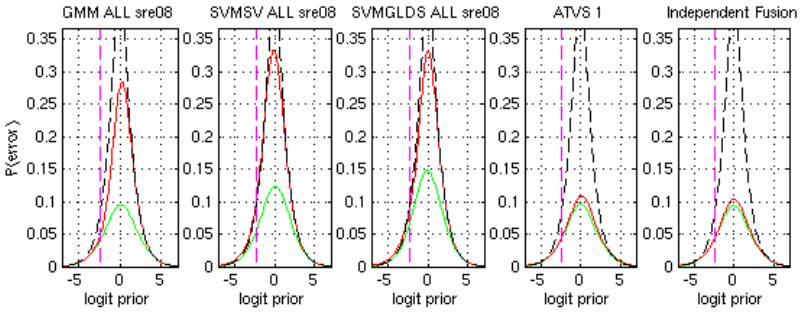
ATVS1@SRE08 across conditions

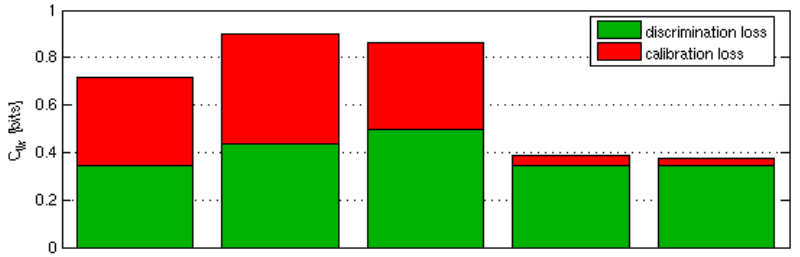


Tested **blindly** over ~100.000 voice comparisons in tel-mic & conv-interview cross conditions

Tippet plot of SRE'08 submitted LRs







The long run towards FASR admissibility

- As calibration is “trained” on known (development) data, systems are “testable” JUST in the assessed conditions
  - Need for caution !!!
- Admissibility is country/court dependent:
  - *Non-Daubert*: case by case
    - Transparent and testable, robust to the mismatch in the case at hand
      - ⇨ channel, session, noise, reverb, duration, language, type of speech, emotional state, …
  - *Daubert*: the technique must be *reliable* (in general)
    - Transparent and testable, robust to mismatch in a wide variety of forensic realistic conditions
    - Challenge:
      - Acceptable error rates & robustness in a variety of mismatched conditions
        - ⇨ Future research: adaptation of NIST-like systems with very limited data to new conditions (variety of scenarios and microphones, car, Lombard, stress …)

The future …

(my vision)

The future of FSR

A good car is nothing without a good pilot !

Perfect coupling between pilot and car is a must !

The feedback from the pilot is critical to improve the car !

A personal tribute …

- **Hermann Künzel - Professor of Phonetics, University of Marburg, Germany**

- From 1985 to 1999, he was Head of the Speaker Identification & Tape Authentication Department of the Federal Criminal Police Office (BKA) in Wiesbaden, Germany.
- He was essential in the development of classical acoustic-phonetic method of forensic speaker recognition (FSR)
- Tutorial on FSR at ESCA Workshop SpkRec (Martigny, 1994)

Last four years: again a pioneer …

Formula One Pilot driving (an automatic system) in more than 100 races (cases) through german, english and turkish circuits (languages)!!!

A message to the students !

Pilots and Mechanical Engineers are welcome !!!

More after coffee break …

INTERSPEECH 2008
SPECIAL SESSION
Forensic Speaker Recognition
Traditional and Automatic Approaches

Format:
The Session will consist of five 20-minute oral presentations, followed by a 20-minute panel discussion.

Plaza 3&4, Time 10:00 – 12:05, Thursday 25th September 2008

References

1. C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.
2. D. J. Balding, *Weight-of-Evidence for Forensic DNA Profiles*, Wiley, 2005.
3. N. Brummer and J. du Preez, “Application independent evaluation of speaker detection”, Computer Speech and Language, vol. 20, no. 2-3, pp. 230-275, 2006.
4. C. Champod and D. Meuwly, “The inference of identity in forensic speaker recognition”, Speech Communication, vol. 31, pp. 193-203, 2000.
5. J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos, M. Garcia-Gomar and J. Ortega-Garcia “Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition”, Computer Speech and Language, vol. 20, no. 2-3, pp. 331-355, 2006.
6. J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano & J. Ortega-Garcia, “Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition”, IEEE Trans. Audio Speech and Language Processing, vol. 15, no. 7, pp. 2104-2115, September 2007.
7. P. Rose, *Forensic Speaker Identification*, Taylor & Francis, 2002.
8. M. J. Saks and J. J. Koehler, “The coming paradigm shift in forensic identification science”, *Science*, vol. 309, no. 5736, pp. 892--895, 2005.

60